УДК: 004.89

ІНФОРМАЦІЙНА СИСТЕМА КЛАСТЕРИЗАЦІЇ ДАНИХ МІСЦЕВИХ ФІНАНСІВ

Коткова А. А.

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Україна, Київ

Роботу присвячено інформаційній системі кластеризації даних місцевих фінансів. Визначено змістовну постановку задачі, описано вхідні, вихідні дані, описано методи розв'язання. Основною метою роботи є покращення розподілу місцевих бюджетів на різні групи шляхом кластеризації даних місцевих фінансів. У Data Mining легше виділити групи схожих об'єктів, у нашому випадку – це схожі місцеві бюджети, вивчити їх особливості та навести аналіз для кожної групи окремо, ніж створювати одну загальну модель для всіх даних.

Ключові слова: місцеві бюджети, доходи, витрати, динаміка бюджету, кластеризація.

Коткова А. А. Комплекс задач кластеризации данных местных финансов / Национальный технический университет Украины "Киевский политехнический институт имени Игоря Сикорского", Украина, Киев

Работа посвящена информационной системе кластеризации данных местных финансов. Определено содержательную постановку задачи, описаны входящие, исходящие данные, описаны методы решения. Основной целью работы является улучшение распределения местных бюджетов на разные группы путем кластеризации данных местных финансов. В Data Mining легче выделить группы похожих объектов, в нашем случае - это похожие местные бюджеты, изучить их особенности и провести анализ для

каждой группы отдельно, чем создавать одну общую модель для всех данных.

Ключевые слова: местные бюджеты, доходы, расходы, динамика бюджета, кластеризация.

A. A. Kotkova Software Solution for clustering local state finance data / National Technical University of Ukraine "Igor Sikorsky Kiev Polytechnic Institute", Ukraine, Kiev

The work is devoted to the information system for clustering local finance data. The substantive formulation of the problem has been defined, incoming, outgoing data, solution methods have been described. The main goal of the work is to improve the distribution of local budgets to different groups by clustering local finance data. In Data Mining, it is easier to identify groups of similar objects, in our case these are similar local budgets, study their features and carry out analysis for each group separately, rather than creating one common model for all data.

Key words: local budgets, revenues, expenses, budget dynamics, clustering.

Вступ. В Україні практично кожного дня зростає кількість відкритих даних, що публікують державні та комерційні органи. На даний момент налічується близько 150 державних реєстрів та дата сетів від державних органів влади. З огляду на реформу децентралізації в Україні аналіз місцевих фінансів є дуже актуальною задачею.

Основна **мета роботи** – покращення розподілу місцевих бюджетів на різні групи шляхом кластеризації даних місцевих фінансів.

Для реалізації поставленої мети необхідно розв’язати такі **задачі**:

1) завантаження даних із порталу openbudget.gov.ua;

2) завантаження даних до бази даних;

3) створення засобів візуалізації даних;

4) створення засобів аналізу даних (кластеризація даних місцевих бюджетів).

Масиви даних можна переглянути на таких порталах: Spending, Відкритий бюджет – бета, Офіційний портал Києва, Відкритий бюджет Херсонської області, Відкритий бюджет Кривого Рогу. У 2018 році було презентовано портал Open Budget, де публікуються дані місцевих бюджетів, на якому є можливість завантажити дані. Аналіз показав, що в аналогів відсутні засоби інтелектуального аналізу цих даних. Наприклад, кластеризація, класифікація. Тому й виникла необхідність реалізувати комплекс задач кластеризації даних місцевих фінансів.

Цільовою аудиторією даної розробки є спеціалісти фінансових органів місцевих бюджетів.

Змістовна постановка задачі. Країна складається з 25 адміністративно-територіальних одиниць. Для кожної з них зібрана інформація по доходах та витратах по всіх населених пунктах. Для ефективного аналізу бюджетів необхідно виконати кластеризацію за показником «Розпис на рік». Задачею кластеризації є розподілення об’єктів на «схожі» групи. Такі групи називають кластерами [1]. Рішенням поставленої задачі є розподілення всіх об’єктів до відповідних кластерів.

Вхідні дані. Усі вхідні дані комплексу задач можна розділити на дві категорії: вхідні документи комплексу задач та вхідні дані користувача. Вхідними документами комплексу задач є рішення сесії про затвердження місцевого бюджету та розпис місцевого бюджету. У

таблиці 1 наведено детальний опис вхідних документів для розпису місцевого бюджету.

Таблиця 1

Опис вхідних документів

Найменування	Реквізити
Розпис доходів місцевих бюджетів	Тип фонду, код, найменування коду, розпис на рік, кошторис на рік, виконано всього, відсоток виконання
Розпис витрат місцевих бюджетів	Тип фонду, КПК, найменування КПК, КФК, найменування КФК, КЕКВ, найменування КЕКВ, розпис на рік, кошторис на рік, виконано всього, відсоток виконання

Вхідними даними користувача є задання кількості кластерів для здійснення кластеризації.

Вихідні дані. Вихідними даними є візуалізовані результати кластеризації на площині декартових координат (рисунок 1).

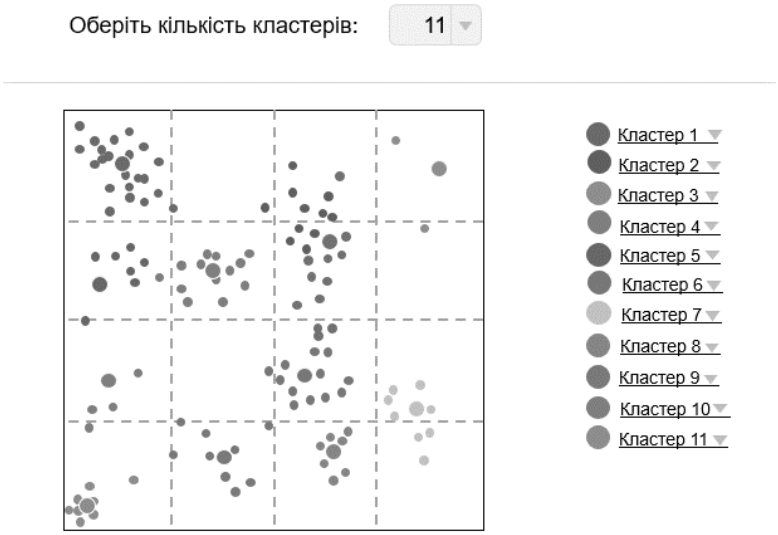


Рис. 1. Візуалізовані результати кластеризації

Завдяки таким результатам можна знайти приховані закономірності в даних. Для кожного кластеру можливо переглянути назви місцевих бюджетів, які в нього увійшли.

Методи розв'язання. Завданням кластеризації для поставленої задачі є розбиття вибірки на кластери, для чого необхідно визначити міру подібності даних [2, с.23–34].

Існують різні способи обчислення міри подібностей [3, с. 112–119]. У роботі було використано евклідову відстань. Формула для розрахунку:

$$\rho(X,Y) = \sqrt{\sum_{i}^{n}(x_i - y_i)^2} \quad \text{, де} \qquad (1)$$

x_i – початкова точка діапазону; y_i – кінцева точка діапазону; n – кількість відрізків, відстань між точками в яких розраховується.

Такий спосіб обчислення міри подібності доцільно приміняти в тому випадку, коли ознаки, що досліджуються, мають однакові величини. Найбільш поширеними алгоритмами кластеризації даних є k-means [4, с.281-288], g-means [5] та c-means [5]. У роботі [6, с.36 – 40] проведено аналіз різних алгоритмів кластеризації та досліджено їх швидкодію. Показано, що алгоритм k-means має високу швидкодію та паралелізацію, тому вирішено використати цей алгоритм.

У роботі було використано ітераційний алгоритм неїєрархічної кластеризації k-means. Для алгоритму k-means розглядаються точки даних та кластери [7]. Алгоритм працює таким чином, що повинен зробити мінімальним сумарне квадратичне відхилення точок кластерів від їх центрів:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 \quad \text{, де} \qquad (2)$$

k – кількість кластерів; x_j – вектори об'єкта кластеризації; S_i – отримані кластери, i = 1, 2, ..., k; μ_i – центри мас векторів $x_j \in S_i$.

Опишемо покрокову реалізацію алгоритму k-means.

КРОК 1. Задаємо, що змінна j рівна 0.

КРОК 2. Випадковим чином ставимо у відповідність кожен елемент початкової множини одному з m кластерів.

КРОК 3. Визначаємо центр кожного кластера. Для нашої задачі це елемент, компоненти якого обчислюються шляхом знаходження середнього арифметичного тих компонентів, які входять у цей кластер. Після цього в центрі кластера досягається мінімум функції суми квадратів відстаней від елементів кластера до точки.

КРОК 4. Обчислюємо відстань від кожного елементу до центру кожного кластеру. Для обчислення міри подібностей використовується Евклідова відстань. Той елемент, для якого ця відстань виявиявилася мінімальною, приписується до кластера. ЯКЩО сума квадратів відстаней від кожного елемента до центру його кластера менша за критичне значення, ТО КРОК 5.

КРОК 5. Вихід. ІНАКШЕ КРОК 6.

КРОК 6. Збільшуємо значення змінної j на 1 та переходимо до КРОКУ 2.

Розглянемо розв'язання поставленої задачі на конкретному прикладі.

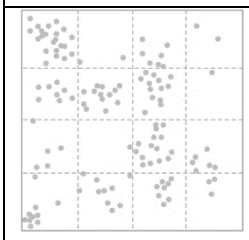
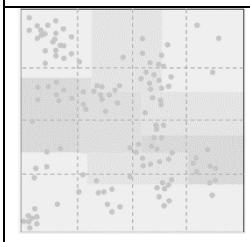
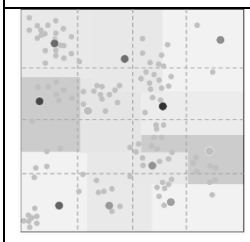
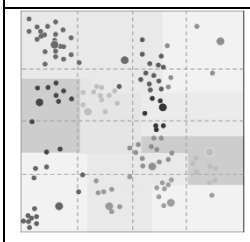
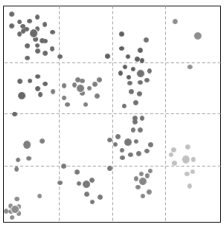
Дано: вхідний документ, у якому міститься інформація про доходи по 126 місцевих бюджетах, кількість кластерів – 11. Рішення нведено в таблиці 2.

Програмна реалізація складається з технологій JS, Node.JS, HTML, CSS, нереляційної бази даних MongoDB.

На рисунку 2 наведено схему нереляційної бази даних.

Таблиця 2

Рішення поставленої задачі

Ілюстрація	Пояснення
	На рисунку наведено 126 місцевих бюджетів, зображених у вигляді точок на області
	Умовно розбиваємо початкову область на 11 кластерів
	Задаємо першочергові точки центрів кластерів. Вибір цих точок відбувається випадковим чином
	Розраховуємо відстані від кожного елемента до центрів кластерів та розприділяємо дані по кластерам залежно від того, на якій відстані вони знаходяться від центра кластера. Якщо відстань від певної точки до центра мінімальна, то точка приписується до того кластера, якому належить центр
	Перераховуємо центри кластерів як центри мас елементів у кластері. Після цього центри кластерів зміщуються та виникає необхідність перерахувати відстані від кожного елемента до центрів мас кажного кластера. Такі розрахунки необхідно проводити до тих пір, доки не буде знайдено стабільне розбиття.

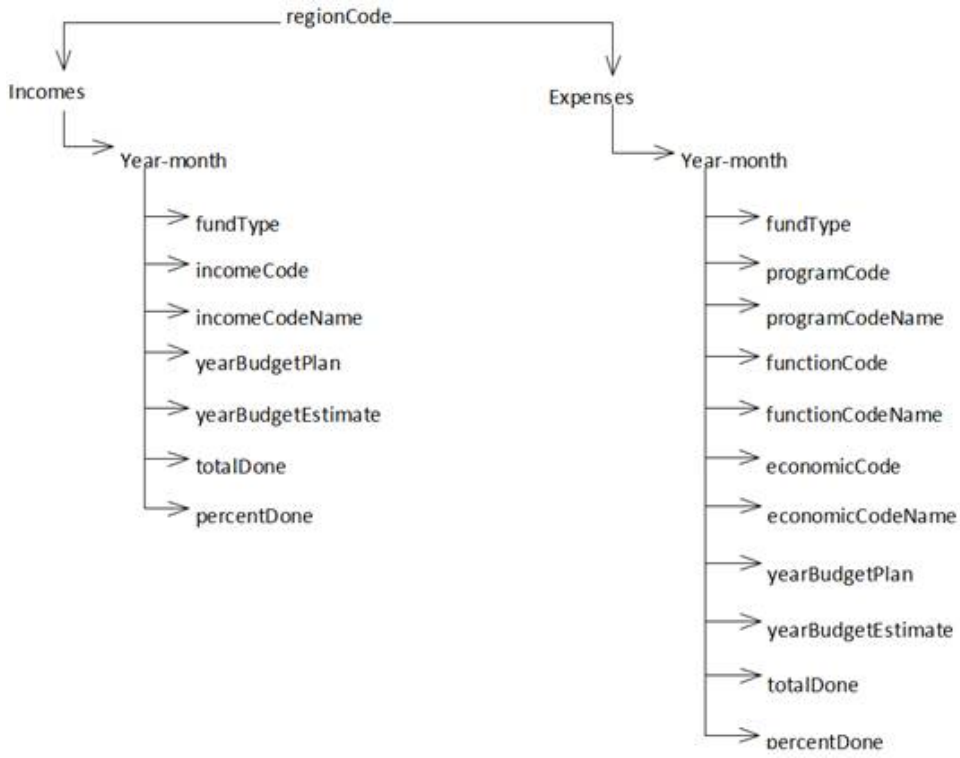


Рис. 2 – Схема бази даних

Висновок. Розроблений програмний продукт є корисним для фахівців, що працюють у сфері економіки, а саме для тих, хто займається питаннями фінансування місцевих бюджетів. Для бюджетів зі «схожими» характеристиками можна підбирати однакові чи подібні програми розвитку.

Було досліджено методи кластеризації k-means, c-means, g-means та обрано k-means. Проведено експримент з кластеризацією на 126 територіальних бюджетах, у результаті чого створено 11 кластерів даних. Перспективою досліджень є застосування інших методів інтелектуального аналізу даних, а саме: класифікації, прогнозування. У комплексі з кластеризацією ці методи дадуть потужний апарат для вирішення серйозних проблем місцевих бюджетів.

Література:

1. Методы и модели анализа данных: OLAP и Data Mining (2009). Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И., СПб., БХВ-Петербург, 512 с.

2. Krzanowski W. J., Lai W. J. (1985). Criterion for determining the number of clusters in a data set. Biometrics, 23–34.

3. Волосюк Ю. В. (2018). Аналіз алгоритмів кластеризації для задачі інтелектуального аналізу даних. Інформаційні технології., 112–119

4. Hamerly G., Elkan C. (2003). Learningthe k in k-means. Whistler, British Columbia, Canada, 281-288

5. Koppen M., Kasabov N., Coghill G. (2008). Advance in Neuro-Information Processing. ICONIP, 342 p.

6. Кіндзерський О.В., Олійник Ю.О. Реалізація алгоритму кластеризації даних k-means на основі технології NVIDIA CUDA. Матеріали наукової конференції студентів, магістрів та аспірантів, Київ, 7, 36 – 40, (23 – 24 квітня 2018 р.)

7. Hartigan J. A. (1975). Clustering Algorithms. Wiley, 369 p.

References:

1. Metody y modely analyza dannykh: OLAP y Data Mining (2009). Barsehian A. A., Kupryianov M. S., Stepanenko V. V., Kholod Y. Y., SPb., BKhV-Peterburh, 512 s.

2. Krzanowski W. J., Lai W. J. (1985). Criterion for determining the number of clusters in a data set. Biometrics, 23–34.

3. Volosiuk Yu. V. (2018). Analiz alhorytmiv klasteryzatsii dlia zadachi intelektualnoho analizu danykh. Informatsiini tekhnolohii., 112–119

4. Hamerly G., Elkan C. (2003). Learningthe k in k-means. Whistler, British Columbia, Canada, 281-288

5. Koppen M., Kasabov N., Coghill G. (2008). Advance in Neuro-Information Processing. ICONIP, 342 p.

6. Kindzerskyi O.V., Oliinyk Yu.O. Realizatsiia alhorytmu klasteryzatsii danykh k-means na osnovi tekhnolohii NVIDIA CUDA. Materialy naukovoi konferentsii studentiv, mahistriv ta aspirantiv, Kyiv, 7, 36 – 40, (23 – 24 kvitnia 2018 r.)

7. Hartigan J. A. (1975). Clustering Algorithms. Wiley, 369 p.